الجامعة المصرية اليابانية للعلوم والتكنولوجيا

E-JUST

Egypt - Japan University of Science and Technology

エジプト日本科学技術大学

CSE 434 - Machine Learning
Final Project Report: Multi-Model Classification of Malignant Lymphoma Subtypes using Transfer Learning and XAI

Shahd Mohamed Saber Ammar - 120220098
Abdallah Adel Shabaan Abdallah - 120220107
Nadine Mostafa Saad AlSayad - 120220310

Submitted to:
Prof. Waleed Gomaa
Eng. Sara Helal

December 23, 2025

1 Introduction

1.1 Problem Statement

In critical medical classification tasks, the predictions produced by machine learning models cannot be accepted without scrutiny, as models may rely on false correlations or irrelevant contextual features rather than clinically meaningful patterns. For this reason, explainable artificial intelligence (XAI) techniques are essential to ensure the reliability and trustworthiness of medical image classification systems.

1.2 Solution Overview

This report presents the implementation of an image classification model for malignant lymphoma detection using a medical imaging dataset. Multiple neural network architectures are evaluated, and the Grad-CAM++ technique is applied to visualize the regions influencing the model's predictions to assess whether classification decisions are driven by tumor-related features or surrounding artifacts.

2 Methodology

2.1 Dataset

This project utilizes the **Malignant Lymphoma Classification** dataset, consisting of 374 high-resolution images (1388 × 1040 pixels) spanning three classes: chronic lymphocytic leukemia (CLL), follicular lymphoma (FL), and mantle cell lymphoma (MCL). The distribution of images across these classes is illustrated in Figure 1.

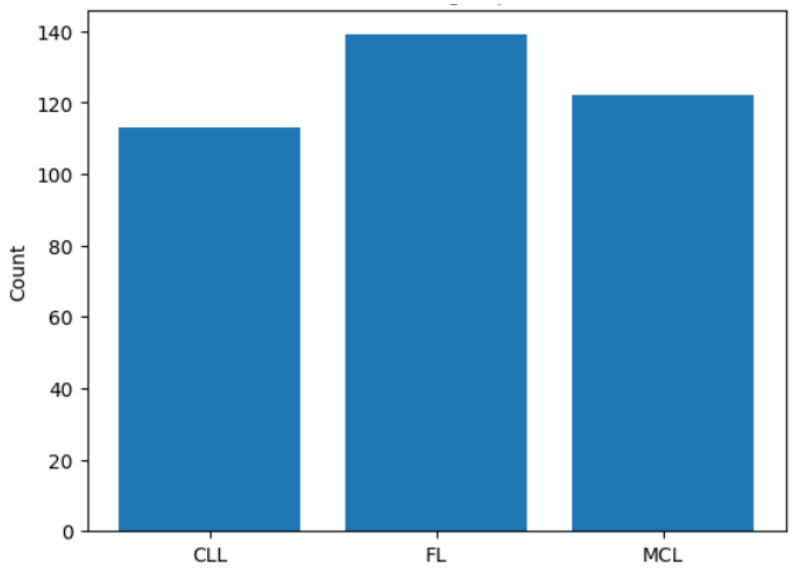


Figure 1: Number of images per class

The data was split into 70% training, 15% validation, and 15% testing sets. All images were resized to 224×224 pixels and normalized using ImageNet-standard mean and standard deviation values. The training set was subjected to a data augmentation pipeline to increase sample diversity, as shown in Table 1.

Table 1: Training Data Augmentation

Transformation	**Value / Range**
Image Rotation	±10°
Shifting	±10% (X and Y)
Shear	±10°
Scaling	0.9× to 1.1×
Horizontal Flip	Left-to-right mirror
Brightness	0.2 to 1.2 intensity
Pixel Filling	Nearest Neighbor

2.2 Model Architectures

This section highlights the different model architectures used in this project.

2.2.1 ResNet50 Architecture

ResNet-50 is a deep convolutional neural network architecture that introduces residual connections to mitigate the vanishing gradient problem and enable the training of very deep models. In this study, a ResNet model pretrained on the ImageNet dataset was used to provide a comparison against DenseNet-based approaches as described later on in this section. ResNet-50 architecture is illustrated in figure 2

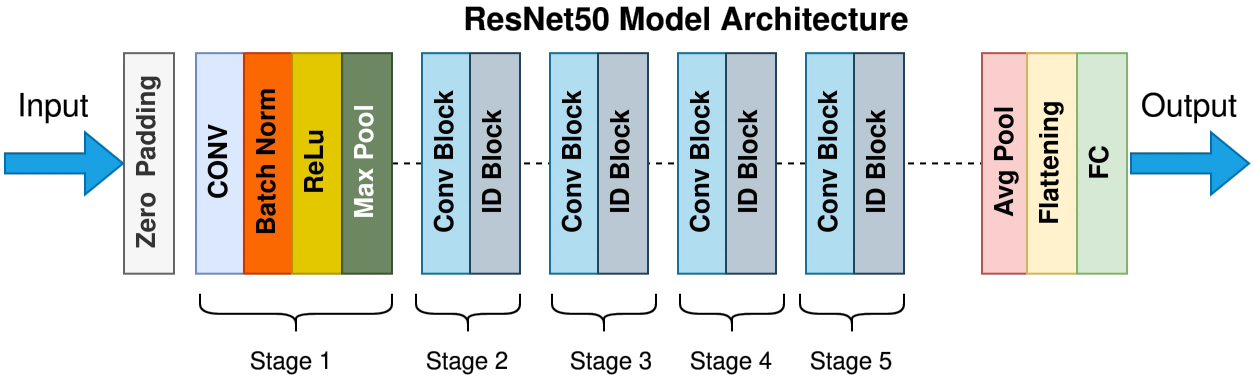


Figure 2: Resnet-50 Architecture

The task of classifying medical biopsy images for lymphoma differs significantly from the natural image classification task on which the ResNet model was originally pretrained using the ImageNet dataset.

Rather than fine-tuning only the final fully connected layer, both the final classification layer and the last convolutional block (layer4, referred to as Stage 5 in Figure 2) were fine-tuned to better adapt the model to domain-specific features present in medical images.

2.2.2 DenseNet-121 Architectures

To provide a robust comparison against ResNet50, two variants of the **DenseNet-121** architecture were introduced. Unlike ResNet, which uses additive identity connections, DenseNet connects each layer to every other layer in a feed-forward fashion within a "Dense Block."

The mathematical representation of a layer l in a DenseBlock is given by:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{1}$$

where $[x_0, x_1, ..., x_{l-1}]$ represents the concatenation of all preceding feature maps. This architecture was selected because its feature-reuse mechanism is highly effective at preserving low-level cellular textures, which are often lost in deeper residual networks.

The intuition behind the DenseBlock can be shown in figure 3

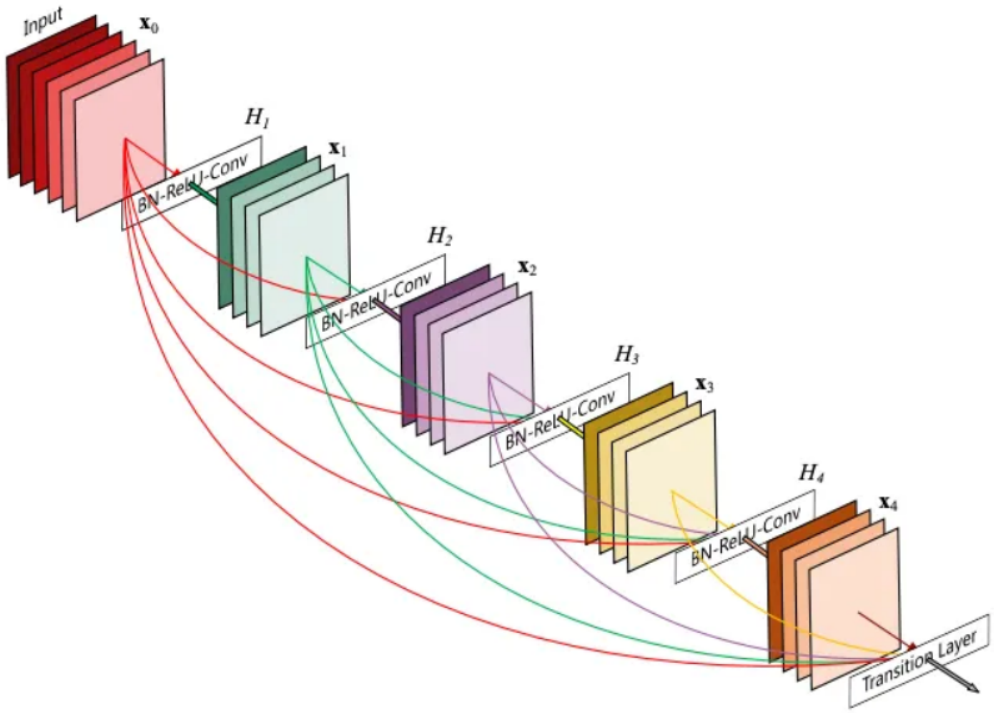


Figure 3: DenseBlock Architecture

DenseNet architecture is based on a series of dense blocks, each containing multiple convolutional layers. Each dense block takes the output of the previous block as input, as well as the

outputs of all the previous blocks. This creates a dense connectivity pattern between all the layers of the network, allowing information to flow more efficiently through the network.

DenseNet-121 (ImageNet Weights)
The first variant utilized the standard **IMAGENET1K_V1** weights. This model serves as a benchmark for general-purpose transfer learning. The original 1000-class Softmax classifier was replaced with a linear layer tailored to the 3 lymphoma subtypes. To adapt the model to the histopathological domain, a "partial unfreezing" strategy was applied: the final dense block (*denseblock4*) was set to be trainable, while the earlier feature-extraction layers remained frozen. This allows the model to leverage general edge-detection filters while optimizing high-level spatial patterns for lymphoma cells.

DenseNet-121 (Medical/KimiaNet Weights)
The second variant utilized domain-specific pre-training via **KimiaNet**. KimiaNet is a specialized version of DenseNet-121 trained on more than 240,000 image patches from the TCGA (The Cancer Genome Atlas) database which consisting of over 15 million pathology patches across 30 different tumor types. The reason for using KimiaNet over ImageNet is based on the *domain gap* between natural images and microscopy:

- **Texture Awareness:** ImageNet models are optimized for macroscopic object shapes. In contrast, KimiaNet is optimized for the stippled chromatin patterns, nuclear-to-cytoplasmic ratios, and pleomorphism found in digital pathology.
- **Stain Invariance:** Medical pre-training often results in weights that are more robust to the variations in Hematoxylin and Eosin (H&E) staining across different labs.

Similar to the ImageNet variant, this model was fine-tuned by unfreezing *denseblock4*, ensuring a fair and controlled comparison between the two pre-training strategies.

2.3 Grad-CAM++

Grad-CAM++ is an explainability method that highlights image regions most important for a network's prediction. For a target class c, Grad-CAM++ generates a class-specific localization map L^c by computing a weighted linear combination of the feature maps A^k from the final convolutional layer. Unlike standard Grad-CAM, which assumes that all pixels in a feature map contribute equally to the final score, Grad-CAM++ introduces pixel-wise importance coefficients, α_{ij}^{kc}. These coefficients are derived from the second and third-order derivatives of the class score Y^c with respect to the feature map activations A_{ij}^k.
The implementation follows a three-step mathematical process:

1. **Pixel-wise Importance Coefficients**:

   The coefficient α_{ij}^{kc} for each location (i, j) in feature map k is calculated using higher-order derivatives to capture precise spatial contributions:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_{a,b} A_{ab}^k \left(\frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3} \right)} \tag{2}$$

2. **Feature Map Weights**:

   The global importance weight w_k^c for each feature map k is computed by taking the weighted sum of the rectified gradients:

$$w_k^c = \sum_{i,j} \alpha_{ij}^{kc} \cdot \text{ReLU}\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{3}$$

3. **Final Localization Map**:

   The class-specific saliency map is generated by combining the feature maps with their respective weights and applying the ReLU activation function to focus only on features with a positive influence:

$$L^c = \text{ReLU}\left(\sum_k w_k^c A^k\right) \tag{4}$$

where Y^c is the score for class c, A_{ij}^k is the activation at position (i, j) in feature map k, and the summation over (a, b) in the denominator of the α equation iterates over the entire spatial dimension of the feature map.

Intuitively, w_k^c weights feature maps based on their influence and curvature, producing sharper heatmaps that highlight regions — such as tumor areas — that drive the model's decision.

3 Results and Discussion

3.1 Models Performance

The models were evaluated using four standard classification metrics: Accuracy, Precision, Recall, and F1-Score. The results summarized in Table 1 demonstrate a clear hierarchy in performance based on architecture and pre-training strategy.

Table 2: Final Performance Metrics on Test Set

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ResNet50 (ImageNet)	0.80	0.82	0.80	0.81
DenseNet121 (ImageNet)	0.82	0.83	0.82	0.82
DenseNet121 (Medical)	**0.84**	**0.85**	**0.84**	**0.84**

3.2 Impact of Architecture: ResNet vs. DenseNet

The results show that DenseNet121 consistently outperformed ResNet50, even when using the same ImageNet initialization (82% vs 80% accuracy). This can be attributed to DenseNet's *feature reuse* mechanism. In histopathology, critical diagnostic features exist at multiple scales—from fine nuclear textures (low-level features) to cellular arrangement patterns (high-level features). DenseNet's dense connections ensure that low-level features are preserved and accessible to deeper layers, which appears more effective for lymphoma classification than ResNet's summation-based residual learning.

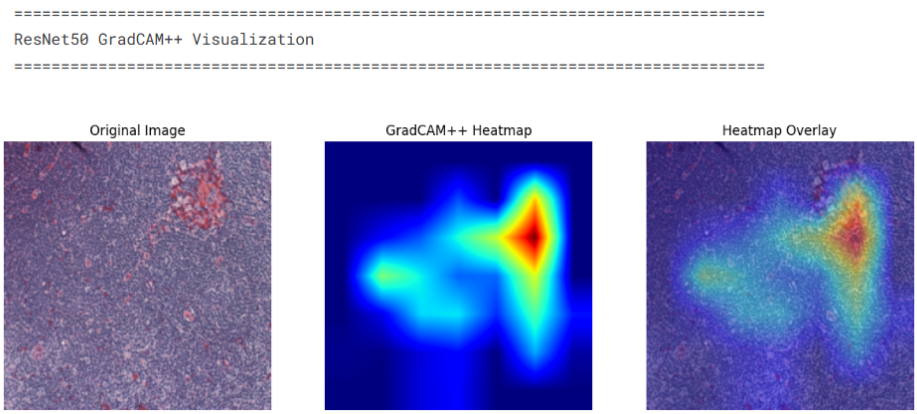
3.2.1 Impact of Domain-Specific Pre-training (KimiaNet)

The **DenseNet Medical** model achieved the highest performance across all metrics (84% Accuracy, 0.84 F1-Score). Unlike the other models pre-trained on natural images (ImageNet), this model was initialized with KimiaNet weights, which were trained on 240,000 pathology patches.

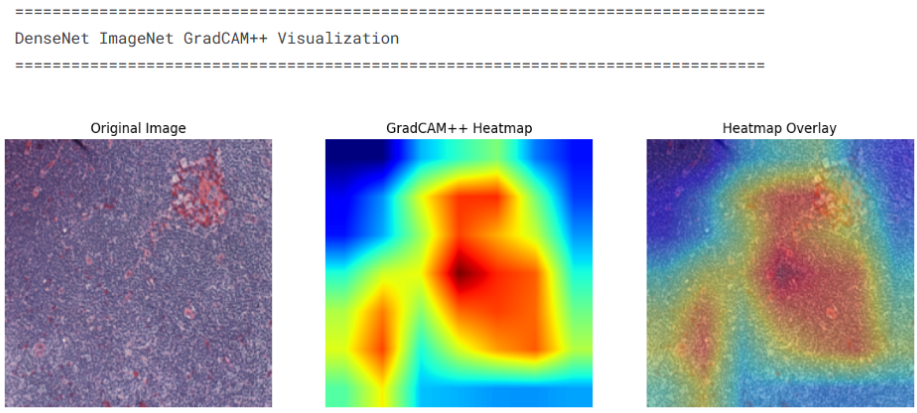
- **Feature Relevance:** ImageNet-based models are trained to recognize edges and shapes of macroscopic objects such as cars, animals, etc. In contrast, the Medical model is already "tuned" to the stippled chromatin patterns and cytoplasmic boundaries of Hematoxylin & Eosin stained slides .
- **Convergence:** The domain-specific starting point allowed the model to reach a more optimal local minimum during fine-tuning, better distinguishing between the small morphological differences of CLL and MCL.

3.3 Explainable AI (XAI) via Grad-CAM++

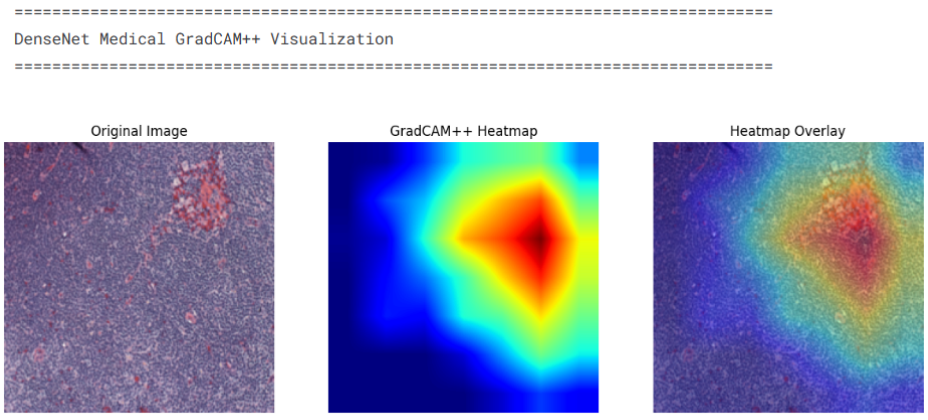
To ensure clinical relevance and validate the diagnostic integrity of the models, Grad-CAM++ was utilized to visualize the decision-making process across all three , as shown in Figure 4. This comparative analysis shows how different pre-training strategies influence which morphological features the models prioritize.



(a) ResNet50



(b) DenseNet (ImageNet)



(c) DenseNet (Medical/KimiaNet)

Figure 4: Comparative Grad-CAM++ visualizations. Each row illustrates how the respective model identifies diagnostic regions within the same CLL tissue sample.

3.3.1 Clinical Interpretability of Heatmaps

The Grad-CAM++ visualizations provided the following insights:

1. **Nuclei Focus:** The "hot" zones (red) are localized primarily over the dense lymphoid nuclei. This indicates the model is correctly identifying the hyper-cellular nature of the lymphoma rather than the connective tissue or empty spaces.

2. **Monotony Detection:** For CLL and MCL, the model's attention is spread across clusters of similar cells. This suggests the network has learned to recognize the "monotonous population" of small lymphocytes, which is a hallmark of these diseases.

3. **Artifact Immunity:** In several test cases, the model successfully ignored red blood cells and staining variations, focusing instead on the nuclear morphology of the malignant cells.

4 Conclusion

This project demonstrates that deep transfer learning is highly effective for lymphoma classification. While general-purpose models like ResNet50 provide high confidence, domain-specific models like DenseNet Medical offer a more better interpretation of pathological slides. The integration of XAI via Grad-CAM++ provides a transparency, verifying that the models are making decisions based on relevant cellular structures, that increases their potential for clinical decision support. Although the current results are promising, expanding the dataset in future work would likely improve models accuracy and generalization by incorporating a more diverse set of training examples.

5 References

Adel, A., Ammar, S., & Elsayed, N. (2024). Xai for lymphoma diaganosis. https://www.kaggle.com/code/shahdammar/xai-for-lymphoma-diaganosis

Larxel. (2020). Malignant lymphoma classification. https://www.kaggle.com/datasets/andrewmvd/malignant-lymphoma-classification

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2017). Grad-cam++: Improved visual explanations for deep convolutional networks. *arXiv preprint arXiv:1710.11063*. https://arxiv.org/abs/1710.11063

Resnet50 — PyTorch documentation. (n.d.). https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html

Densenet121 — PyTorch documentation. (n.d.). https://docs.pytorch.org/vision/main/models/generated/torchvision.models.densenet121.html

Kimia Lab Mayo. (2021). KimiaNet: A trained network for histopathology image representation.